Sur un réseau de bibliothèques hypertextes, un atelier flexible d'exploration de corpus en sciences humaines et expérimentales

WicriExplore

Introduction

De la musicologie à l'IST en passant par la santé

Dilib / Wicri / LorExplor / WicriExplore

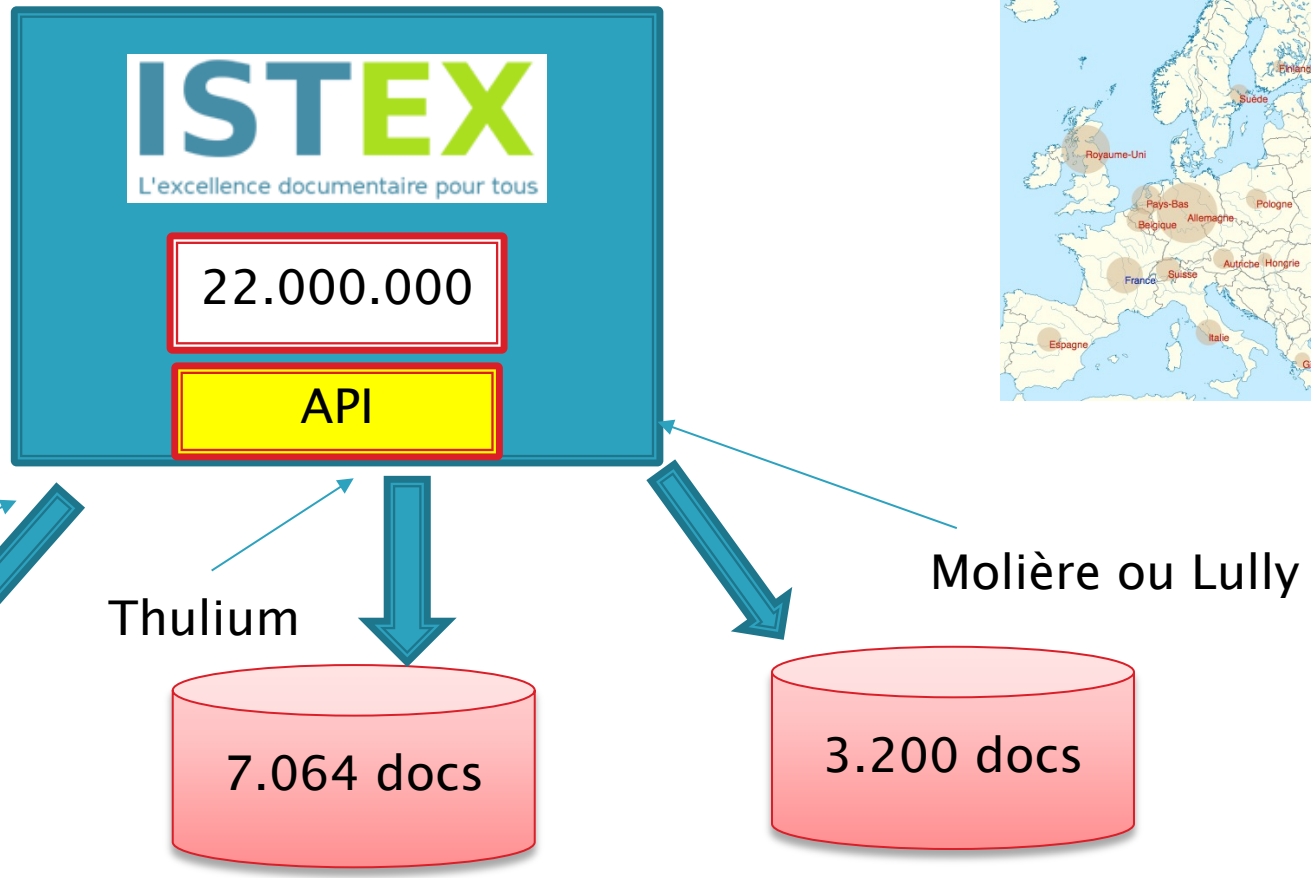
- Dilib (INIST / Loria 1992 – 2002)
  - Une **boîte à outils XML** *(Document & Information Library)*
  - Un lego pour construire un système de recherche d'information
- Wicri (Univ. Lorraine 2008 – …)
  - Un réseau de **wikis sémantiques**
  - WIkis des Communautés de la Recherche et de l'Innovation
- LorExplor (UL 2014 + ISTEX – …)
  - Exploration des besoins lorrains avec ISTEX (22.000.000 d'articles)
- Et maintenant WicriExplore (Paragraphe 2020)
  - Applications de l'ensemble avec un point fort
    - dans les humanités (musicologie, Chanson de Roland, histoire de l'IST)
  - Sans oublier les sciences expérimentales (santé, environnement…)

ISTEX

L'excellence documentaire pour tous

- Projet français
  - (Investissements d'avenir)
- Slogan
  - *L'excellence documentaire pour tous*
- Equipement
  - Budget : environ 60.000.000 €
    - 50.000.000 € Archives et données
      - Springer, Wiley, Elsevier, Oxford University Press… (texte ou XML…)
    - Un portail opérationnel : 6.000.000 €
    - Soutien recherche : 1.000.000 € (une dizaine d'expériences)
    - LorExplor : 100.000 € (150 explorations de corpus)

ISTEX
des millions de documents à explorer



En santé : volumétrie débordante

- Covid
  - Google : 8 210 000 000 pages
  - Google Scolar : 4 170 000 pages ou articles scientifiques
  - PubMed : 410 731
  - PubMed Central : 656 356
  - HAL : 15 992
  - Wikipédia (fr) : 25 900
  - Wikipedia (en) 126 656
- Sur WicriExplore , thématique des épidémies grippales
  - 24 serveurs d'exploration
  - 70.000 documents

Le Covid avec WicriExplore (en 2021)

	HAL	ISTEX	PubMed	PMC	Pascal	Francis	Total	Total corrigé
Serveur d'exploration Chloroquine	262	4 500	701	1 125	104	0	6 692	6 435
Serveur d'exploration Covid	13	1 052	767	814	71	1	2 718	2 451
Serveur d'exploration Covid (26 mars)	13	1 052	1 421	1 144	71	1	3 702	3 214
Serveur d'exploration H2N2	340	2 508	665	1 454	140	0	5 107	4 660
Serveur d'exploration MERS	25	4 000	4 326	2 049	118	1	10 519	8 450
Serveur d'exploration SRAS	203	3 707	5 359	2 563	989	6	12 827	10 380
Serveur d'exploration Stress et Covid	224	3 573	1 145	1 000	294	88	6 324	5 969
Serveur d'exploration Tocilizumab	86	2 264	3 019	1 825	263	1	7 459	6 869
Serveur d'exploration sur la grippe au Canada			1 386				1 386	1 386
Serveur d'exploration sur la grippe en Allemagne			634				634	634
Serveur d'exploration sur la grippe en Belgique			130				130	130
Serveur d'exploration sur les pandémies grippales	837	3 967	3 311	1 574	3 046	113	12 849	10 482

Dilib, une histoire

- 1973 Nancy Centre de calcul (IUCAL)
  - Le TLF (Trésor de la langue française)
  - Le logiciel Mistral de la Cii
- 1980 GS ANL
  - Base des logiciels IA + GL
  - La SM90 machine Unix française
- 1988 INIST – LORIA – INIST
  - 1990 : Ilib : SGML pour Pascal et Francis
  - 1992 : DILIB : Exploration XML de corpus hétérogènes
- 2002 : Colloques et revues (CIDE … Artist)
- 2008 : WICRI
- 2012 : LorExplor : Wicri + Dilib
  - traiter du corpus volumineux,
    - Textuel, multi-dtd
  - Réseau MediaWiki
    - Générations de modèles wiki
    - Robots

Exploration, Filtrage

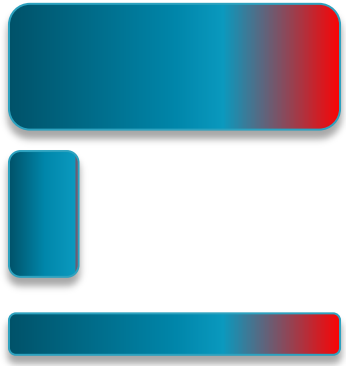
- Quelles sont les œuvres de Mozart les plus citées dans un corpus ?
  - Idée générale : utiliser le catalogue Köchel
    - Résultat : Sonate KV. 448

```
HfdCat Data/Main/Exploration/biblio.hfd                \
  | SxmlFindText -r "[K][Vv]*[ \.]*[0-9][0-9]* »  \
  | SxmlSelect -p @5 -p @1 | sort | IndexBuildRec
```

- Quelles sont les applications de « *dance therapy* » avec une dimension artistique ?
  - Recherche de présence de chorégraphes (nom-prénom) en utilisant un filtre créé pour les noms binomiaux

Corpus : méfiance / Hallucinations/ curation

- Exemple : Mozart
  - 15.000 documents (Musique + médecine)
  - Quelques problèmes de type « avenue Mozart »
  - Plus sérieux :
    - Musique : peu de signalement d'affiliations
    - Médecine : forte politique d'affiliations
  - Les statistiques se focalisent sur la médecine…
- Exemple : Parkinson en France
  - Parkinson : 90.000 documents
  - Extrait de 4000 documents :
    - peu de bruit
  - Parkinson en France :
    - beaucoup de bruit.



- Les perturbations potentielles de l'OCR
  - Serveur sur la méthode Scrum pollué à 85% par des corrections OCR
  - sérum -> scrum